Language Diarization for Indian Languages

EE6307
Speech Systems

Venkatesh Parvathala
Akhil Kumar Donka
Indian Institute of Technology Hyderabad

December 16, 2022

Datasets

- The code-switching data for Indian languages is not readily available !!
  - WSTCSMC has three code-switching pairs : Gujarati-English, Hindi-English and Telugu-English but this dataset is not open sourced.
  - MUCS 2021 has two code-switching pairs [1] : Hindi-English and Bengali-English but we were not able to download the data.
- Other possibilities ?
  - **Creating Dataset by ourselves** - It requires a lot of human effort and therefore we are not interested in this.
  - **Using monolingual data** - we can get monolingual data for all Indian languages but most of them will also contain code-switches.
    - **Synthetic data using monolingual data?**
- In this project, we would like to use monolingual data.

[1] `https://navana-tech.github.io/MUCS2021/data.html`

Method-1 to leverage monolingual data

- Estimate the PDF of features for each language using GMM.
- To test the language identification performance, extract the features for the test utterance and assign the utterance with the maximum likelihood language label

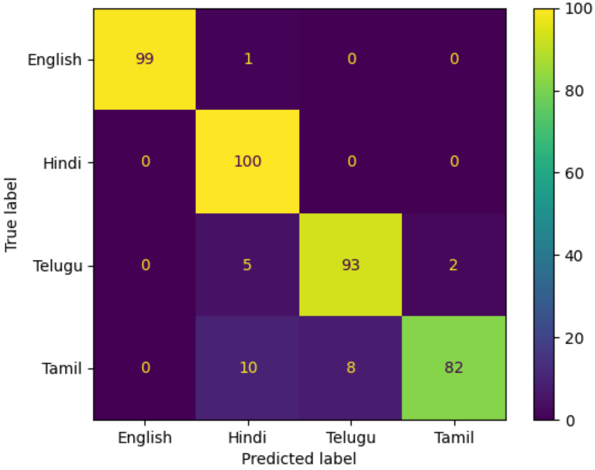
$$L = arg\ max_k p(X/\lambda_k) \tag{1}$$

- We considered four Indian languages - Telugu, Tamil, Hindi and English (Indian English).
- Telugu, Tamil and Hindi data was taken from Kaggle[2] and the English dataset is taken from IITM ASR challenge.
- We have taken 7 hours of data from each language.
- 39 dimensional mfccs are extracted and the 64 mixture GMM is trained.

[2] https://www.kaggle.com/datasets/hbchaitanyabharadwaj/audio-dataset-with-10-indian-languages

Confusion matrix

- Accuracy : 93.5 %



True label \ Predicted label	English	Hindi	Telugu	Tamil
English	99	1	0	0
Hindi	0	100	0	0
Telugu	0	5	93	2
Tamil	0	10	8	82

GMM for diarization

- For a given test utterance, divide the utterance into small segments(around 1 to 2sec) and assign the segment with the maximum likelihood language label

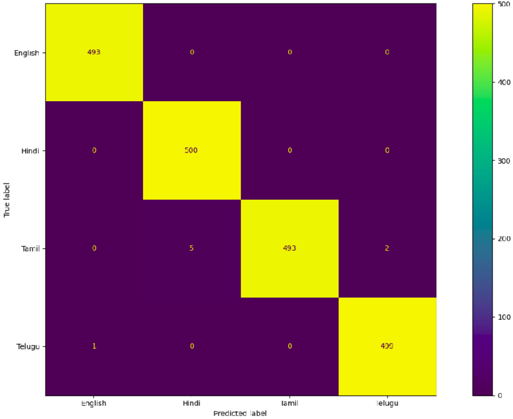
$$L_t = arg\ max_k p(X_t/\lambda_k) \tag{2}$$

- After getting the predictions for each segment, use BIC as post processing step to filter the irregularities.
  - Can we replace the mfcc with the extracted language embeddings from a trained language identification system for BIC ?

Method-2 to leverage monolingual data

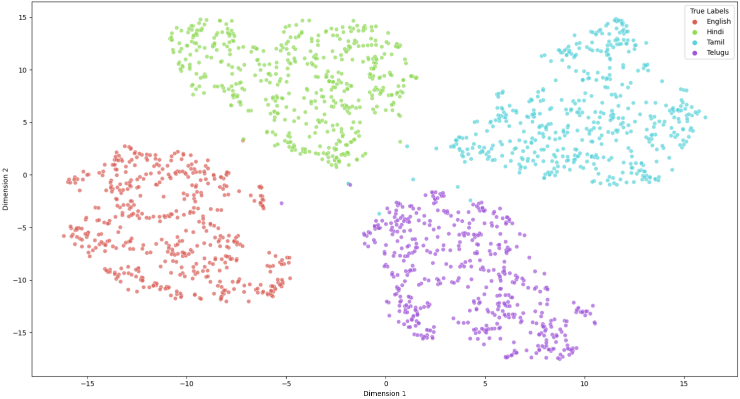
- Use the BIC or any segmentation/clustering method as a preprocessing step to segment the utterance with the switches.
- Now use a speaker identification system to identify the language of each segment.
- The speaker identification system can be an x-vector network trained on the available LID data.
- We trained an x-vector network on the same data that was used for GMM.
- Features : 64 Log mel features with 25ms window & 10ms hop length
- We obtained 99% accuracy in this case.

Confusion matrix



True label \ Predicted label	English	Hindi	Tamil	Telugu
English	493	0	0	0
Hindi	0	500	0	0
Tamil	0	5	493	2
Telugu	1	0	0	499

t-SNE plot


True Labels
● English
● Hindi
● Tamil
● Telugu
Dimension 1
Dimension 2
15
10
5
0
−5
−10
−15
−15
−10
−5
0
5
10
15

Future directions and Remarks

- We observed that both the LID systems are overfitting to the data.
- More diversified data might be required to train the networks (with more number of speakers).
- Using LID model on the segments is one straight forward technique.
- Clustering techniques to be explored as a preprocessing step.